Homework #2

See Canvas for HW #2 assignment due date.

A. Theoretical Problems

Problem A.1

Keep in mind that during class, we kept our equations in the 1,..,p-1 space for predictors. We will use this HW to get familiar with the other very common way of doing it.

Matrices and vectors will play an important role for us in linear regression. Let's review some matrix theory as it might relate to linear regression.

Consider the system of linear equations

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} + \varepsilon_i,$$

for $i = 1, \ldots, n$, where n is the number of data points (measurements in the sample), and $j = 1, \ldots, p$, where

1. $p + 1$ is the number of parameters in the model.
2. Y_i is the i^{th} measurement of the *response variable*.
3. $X_{i,j}$ is the i^{th} measurement of the j^{th} *predictor variable*.
4. ε_i is the i^{th} *error term* and is a random variable (often assumed to be $N(0, \sigma^2)$).
5. β_j are *unknown parameters* of the model, ($j = 0, \ldots, p$). We hope to estimate these, which would help us characterize the relationship between the predictors and response.

(a) Write the equation above in matrix vector form. Call the matrix including the predictors X, the vector of Y_is $\mathbf{Y}$, the vector of parameters β, and the vector of error terms ε. (This is more LaTeX practice than anything else...)

$$\mathbf{Y} = X\beta + \varepsilon,$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $X = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \ldots & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & \ldots & X_{2,p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \ldots & X_{n,p} \end{bmatrix}$, , $\beta = (\beta_1, \ldots, \beta_n)^T$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$

(b) In class, we will find that the OLS estimator for β in MLR is $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y}$.

1. What condition must be true about the columns of X for the "Gram" matrix $X^T X$ to be invertible?
2. What does this condition mean in practical terms?

3. Suppose that the number of measurements (n) is less than the number of model parameters ($p+1$). What does this say about the invertibility of X^TX? What does this mean on a practical level?
4. What is true about about $\widehat{\boldsymbol{\beta}}$ if X^TX is not invertible?

1. For X^TX to be invertible, the columns of X must be linearly independent. That means that no column of X--i.e., no measured predictor--can be written as a linear combination of other columns. This implies that $n > (p+1)$.
2. If we've measured a predictor that is simply a linear combination of others, that means that that predictor is not adding any new information that's not already contained in the other predictors. Imagine a simple case: X_1 is a predictor of measured weights in pounds, and X_2 is a predictor of measured weights in kilograms. Thus, $X_1 = 2.2046X_2$. Measuring X_1 doesn't give any new information.
3. This implies that we have more parameters than data/measurements, and thus X^TX will not be invertible.
4. The formula for $\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{Y}$ is derived from the "normal equations":

$$(X^TX)\beta = X^T\mathbf{Y}.$$

The normal equations have a unique solutions if and only if X^TX is invertible. If it's not invertible, then either the normal equations have no solutions or infinitely many solutions.

Problem A.2

In class, we defined the *hat* or *projection* matrix as

$$H = X(X^TX)^{-1}X^T.$$

The goal of this question is to use the hat matrix to prove that the fitted values, $\widehat{\mathbf{Y}}$, and the residuals, $\widehat{\boldsymbol{\varepsilon}}$, are uncorrelated. We will do it in steps, and *some* of the proofs will only be required for STAT 5010 students. Note that STAT 4010 students are asked to answer part (e), as to why this result has practical importance.

(a) Show that $\widehat{Y} = HY$. That is, H "puts a hat on" Y.

$\widehat{Y} = X\widehat{\boldsymbol{\beta}} = X(X^TX)^{-1}X^T\mathbf{Y} = HY$

(b) Show that H is symmetric: $H = H^T$.

$H^T = \left(X(X^TX)^{-1}X^T\right)^T = (X^T)^T\left[(X^TX)^{-1}\right]^TX^T = X\left[(X^TX)^T\right]^{-1}X^T = X(X^TX)^{-1}X^T = H$

(c) (STAT 5010 Only) Show that $H(I_n - H) = 0_n$, where 0_n is the zero matrix of size $n \times n$.

Note that

$$HH = \left(X(X^TX)^{-1}X^T\right)\left(X(X^TX)^{-1}X^T\right) = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H,$$

because $(X^TX)^{-1}X^TX = I_{p+1}$.

Thus, $H(I_n - H) = HI_n - HH = H - H = 0_n$

(d) (STAT 5010 Only) Stating that $\widehat{\mathbf{Y}}$ is uncorrelated with $\widehat{\boldsymbol{\varepsilon}}$ is equivalent to showing that these vectors are orthogonal.* That is, we need to show that their dot product is zero:

$$\widehat{\mathbf{Y}}^T \widehat{\boldsymbol{\varepsilon}} = 0.$$

Prove this result.

*It's intersting to think about why uncorrelated, in this case, is equivalent to orthogonal. Extra credit if you can tell me why!

$$\widehat{\mathbf{Y}}^T \widehat{\boldsymbol{\varepsilon}} = (H\mathbf{Y})^T (I - H)\mathbf{Y} = \mathbf{Y}^T H^T (I_n - H) Y = \mathbf{Y}^T H (I_n - H) Y = \mathbf{Y}^T 0_n Y = 0.$$

(e) Why is this result important in the practical use of linear regression?

This result shows that, if the linear regression assumptions are met, then there should be no correlation between the fitted values, $\widehat{\mathbf{Y}}$, and the residuals, $\widehat{\boldsymbol{\varepsilon}}$. Thus, it may give us a way to check some of our model assumptions. For example, if we plot the residuals against the fitted values and see a trend, we might conclude that some model assumption is incorrect (we will learn more about "regression diagnostics" in Unit #4).

B. Computational Problems

Problem B.1

Let $X_1, \ldots, X_{30} \overset{iid}{\sim} N(1, 9)$. The formula for a 90% confidence interval for μ is

$$\bar{X} \pm 1.64 \frac{\sigma}{\sqrt{n}}.$$

Let's conduct a simulation to confirm the coverage of this confidence interval.

(a) Generate $m = 500$ random samples of size $n = 30$ from $N(1, 9)$ and calculate the 90% confidence interval for each. Don't print anything.

```
set.seed(0019)
n = 30; mu = 1; sig = 3; m = 500; x = replicate(m, rnorm(n,mu,sig))
xbar = colMeans(x);
ci = cbind(xbar-1.64*3/sqrt(n), xbar+1.64*3/sqrt(n));
```

(b) Estimate the coverage by finding the number of intervals that cover the true mean, and dividing my m.

```
coverage = 1-sum(ci[,1] > mu | ci[,2] < mu)/m
cat("The estimated coverage is",coverage,".")
```

Problem B.2

(a) Load the "gala" dataset, and describe the variables.

```
gala = read.table("https://www.colorado.edu/amath/sites/default/files/attached-files/gala.txt", sep =
```

(b) Use ggplot() to explore the relationship between the Species variable (response) and Endemics, Elevation, Nearest, and Adjacent (predictor variables). You might do so by creating four separate scatter plots. Do these relationships look linear? Does the variability in Species change as a function of any of the predictors? Are there any outliers in any of the plots?

```
#install.packages("gridExtra")
library(gridExtra)
library(ggplot2)

p = ggplot(data = gala)
p1 = p + geom_point(mapping = aes(x = Endemics, y = Species))
p2 = p + geom_point(mapping = aes(x = Elevation, y = Species))
p3 = p + geom_point(mapping = aes(x = Nearest, y = Species))
p4 = p + geom_point(mapping = aes(x = Adjacent, y = Species))
grid.arrange(p1,p2,p3,p4, nrow = 2)
cor(gala)
```

1. The Species vs Endemics plot appears to be approximately linear, but perhaps with some curvature.
2. The Species vs Elevation plot appears roughly linear, but there appears to be more variance in Species as Elevation gets higher.
3. The Species vs Nearest plot doesn't appear to have a linear relationship.
4. The Species vs Adjacent plot doesn't appear to have a linear relationship, and there seem to be a few very large outliers.

(c) Perform a linear regression with Species as the response and Endemics, Elevation, Nearest, and Adjacent as predictors. Interpret the parameter estimate associated with Endemics (assume, for the moment, that the model is correct, and so the interpretation holds).

```
lmod = lm(Species ~ Endemics + Elevation + Nearest + Adjacent, data = gala)
summary(lmod)
```

After adjusting for the highest elevation of the island (m), the distance from the nearest island (km), and the area of the adjacent island (square km), if the number of endemic species on that island were increased by one, *on average*, we would see a ≈ 4.19 increase in the number of plant species found on the island.

(This conclusion sounds causal; but it shouldn't necessarily be interpreted as such!)

(d) Calculate the residual sum of squares, and the total sum of squares for this model. Then, use these calculations to verify the Multiple R-squared calculation in the summary from the previous part. Interpret R^2 (assume, for the moment, that the model is correct, and so the interpretation holds).

```
rss = sum(resid(lmod)^2)
tss = with(gala, sum((Species - mean(Species))^2))
cat("The TSS is ", tss, ".")
cat("The RSS is ", rss, ".")

anova(lmod)
```

Note that the anova table calculates sums of squares. Where is the TSS??

(e) Plot the residuals vs the fitted values. Based on what we've discussed in class (and a question from Section A of this homework!), what do you expect to see in this plot? Do you see what you expect to see? If not, what does that mean?

```
plot(lmod)
```

We would expect to see no correlation between the residuals and the fitted values (the first plot above). But there does seem to be a downward linear trend, for small fitted values (and other issues that we'll discuss in unit #4).

Problem B.3 (STAT 5010 Only)

Here's a procedure for calculating a two-sample bootstrap hypothesis test. You will apply this procedure on real data below.

Let $X_1, \ldots, X_{n_1}$ be an iid sample from population #1, with unknown mean μ_1 and known standard deviation σ_1, and let $Y_1, \ldots, Y_{n_2}$ be an iid sample from population #2, with unknown mean μ_2 and known standard deviation σ_2. Suppose we want to conduct a hypothesis test of the sort:

$$H_0 : \mu_1 - \mu_2 = 0 \ \ vs \ \ H_1 : \mu_1 - \mu_2 \geq 0.$$

The following algorithm has been suggested for a bootstrap test.

1. Calculate the test statistic

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

2. Let $\bar{z}$ be the mean of the combined data sets. Create two new data sets, $x'_1, \ldots, x'_{n_1}$ and $y'_1, \ldots, y'_{n_2}$ that are the original data sets centered at $\bar{z}$.
3. Draw B random bootstrap samples of size n_1 from $x'_1, \ldots, x'_{n_1}$ and of size n_2 from $y'_1, \ldots, y'_{n_2}$. The result will be two matrices, x^* and y^*; x^* will containin columns of bootstrap samples from sample #1, and y^* will contain columns of

bootstrap samples from sample #2.

4. Then, for each bootstrap sample pair, calculate

$$t^* = \frac{\bar{x^*}_j - \bar{y^*}_j}{\sqrt{\sigma_1^{*^2}/n_1 + \sigma_2^{*^2}/n_2}},$$

where $\bar{x^*}_j$ is the sample mean of the j^{th} bootstrap sample from sample #1, and $\bar{y^*}_j$ is the sample mean of the j^{th} bootstrap sample from sample #2. $\sigma_1^{*^2}$ and $\sigma_2^{*^2}$ are the corresponding variance estimates of the j^{th} bootstrap sample. t^* will be a vector of length B and will approximate the distribution of the test statistic t.

5. Estimate the p-value using

$$\frac{\#\text{ of times } \{t^* \geq t\}}{B}.$$

Appling this procedure to real data...

A tennis club has two systems to measure the speed of a tennis ball. The local tennis pros suspect one system, `speed1`, is consistently recording faster speeds. To test her suspicions, she sets up both systems and records the speed of 12 serves (three serves from each side of the court). The values are stored in the data frame `tennis`, with variables `speed1` and `speed2`. The recorded speeds are in kilometers per hour.

Does the evidence support the tennis pro's suspicion? Use the above bootstrap hypothesis testing procedure and $\alpha = 0.1$.

```
tennis = read.table("https://www.colorado.edu/amath/sites/default/files/attached-files/tennis.txt", s
with(tennis, t.test(speed1, speed2, alternative = "greater"))
x = tennis$speed1; n1 = length(x); sig1 = sd(x);
y = tennis$speed2; n2 = length(y); sig2 = sd(y);

t = (mean(x) - mean(y))/(sqrt(sig1^2/n1 + sig2^2/n2)); t

zbar = mean(c(x,y)); xprime = x - mean(x) + zbar;
yprime = y - mean(y) + zbar

B = 500;
xstar = replicate(B, sample(xprime,n1,replace = TRUE))
ystar = replicate(B, sample(yprime,n2,replace = TRUE))

tstar = (colMeans(xstar) - colMeans(ystar))/(sqrt(sig1^2/n1 + sig2^2/n2));


pvalue = sum(tstar >= t)/B; pvalue
```

The p-value is high, and thus we don't have good evidence to support the tennis pro's suspicion.